LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-235865

Chemometric and Statistical Analyses of ToF-SIMS Spectra of Increasingly Complex Biological Samples

E. S. Berman, L. Wu, S. L. Fortson, D. O. Nelson, K. S. Kulp, K. J. Wu

October 29, 2007

Surface and Interface Analysis

Disclaimer

Chemometric and Statistical Analyses of ToF-SIMS Spectra of Increasingly Complex Biological Samples

Elena S.F. Berman, Ligang Wu, Susan L. Fortson, David O. Nelson, Kristen S. Kulp, Kuang Jen Wu

Chemistry, Materials, Earth and Life Sciences Directorate,

Lawrence Livermore National Laboratory,

Livermore, CA 94550

Address correspondence to E. Berman, Lawrence Livermore National Laboratory, 7000 East Ave, L-452, Livermore, CA 94550. Email: berman2@llnl.gov

Abstract

Characterizing and classifying molecular variation within biological samples is critical for determining fundamental mechanisms of biological processes that will lead to new insights including improved disease understanding. Towards these ends, time-of-flight secondary ion mass spectrometry (ToF-SIMS) was used to examine increasingly complex samples of biological relevance, including monosaccharide isomers, pure proteins, complex protein mixtures, and mouse embryo tissues. The complex mass spectral data sets produced were analyzed using five common statistical and chemometric multivariate analysis techniques: principal component analysis (PCA), linear discriminant analysis (LDA), partial least squares discriminant analysis (PLSDA), soft independent modeling of class analogy (SIMCA), and decision tree analysis by recursive partitioning. PCA was found to be a valuable first step in multivariate analysis, providing insight both into the relative groupings of samples and into the molecular basis for those groupings. For the monosaccharides, pure proteins and protein mixture samples, all of LDA, PLSDA, and SIMCA were found to produce excellent classification given a sufficient number of compound variables calculated. For the mouse embryo tissues, however, SIMCA did not produce as accurate a classification. The decision tree analysis was found to be the least successful for all the data sets, providing neither as accurate a classification nor chemical insight for any of the tested samples. Based on these results we conclude that as the complexity of the sample increases, so must the sophistication of the multivariate technique used to classify the samples. PCA is a preferred first step for understanding ToF-SIMS data that can be followed by either LDA or PLSDA for effective classification analysis. This study demonstrates the strength of ToF-SIMS combined with multivariate statistical and chemometric techniques to classify increasingly complex biological samples. Applying these techniques to information-rich mass

spectral data sets opens the possibilities for new applications including classification of subtly different biological samples that may provide insights into cellular processes, disease progress, and disease diagnosis.

Introduction

Biological systems are exceptionally complex, with perturbations of an organism most often evidenced as changes in multiple biological pathways. For this reason, it is increasingly apparent that molecular patterns or "fingerprints," which capture significantly more biological information, may be more useful than single-molecule markers for detecting, classifying and understanding biological changes. Analytical techniques that probe multiple biological molecules of interest, such as mass spectrometry or DNA or protein arrays, are thus finding increasing application to biological questions, especially the difficult problem of informed disease diagnosis and prognosis. Of particular interest is the ability to assign an unknown sample to a known class by comparing the molecular patterns in the unknown with molecular signatures of the known group. This approach for gene array analysis is being used extensively in cancer biology as a means to classify tumors according to cancer phenotype, as reviewed by Brentani et al.[1] and Macgregor. [2]

We are developing methods to use the information-rich spectral data generated by time-of-flight secondary ion mass spectrometry (ToF-SIMS) to classify biological samples. ToF-SIMS is a surface-sensitive mass spectral analysis technique used to detect and localize chemical and molecular information from sample surfaces. ToF-SIMS instruments use a finely focused (optimal spot size of 150 nm) energetic primary ion beam to desorb secondary molecular and fragment ions into a time-of-flight mass spectrometer. These ions can be recorded either as a single spectrum from a sample surface, or as a mass spectral image where each pixel is a

complete mass spectrum. These mass spectral images can then be analyzed as molecular images of single peaks or as an average mass spectrum of a defined area.

Although used routinely for inorganic sample analysis, ToF-SIMS is gaining increasing popularity for the analysis of biomaterials and biological samples.[3-6] The analysis of biological samples presents unique challenges to ToF-SIMS as these samples are at the same time enormously complex and quite similar to one another. Additionally, important constituents of the sample are frequently present at extremely low concentrations. The complexity of biological samples creates complex mass spectra that are difficult if not impossible to interpret by visual inspection. Multivariate analysis and pattern recognition techniques are widely employed for interpretation of such data sets, as they reduce the data complexity and illuminate distinguishing features from the data.

A good deal of recent work has focused on the use of multivariate analysis techniques on ToF-SIMS data of biological samples. We have shown classification of monsaccharides by PCA and LDA[7] and our group as well as others have reported multivariate analysis of pure proteins adsorbed to surfaces and of protein mixtures,[8] reviewed by Michel et al.[5] Jungnickel and coworkers have used PCA to discriminate yeast strains,[9] the Vickerman group has discriminated different types of bacteria[10, 11] and our group has distinguished between human breast cancer cell lines[8] and mouse embryo tissues.[12] Many more recent examples have focused on different types of multivariate analyses, especially imaging analysis, rather than on a specific biological question of interest.[13-15]

This work extends previously published analyses of ToF-SIMS data by analyzing more complex and biologically relevant samples and by comparing a wider variety of multivariate methods for spectral analysis. This study focuses specifically on the task of classification of

samples based on the analysis of a training set of spectra with known groupings. Traditionally, such classification has been accomplished by different methods depending on whether the analyst is a chemist or statistician by training. Chemists tend to use a subset of mathematical and statistical techniques commonly referred to as "chemometrics," which were developed specifically by chemists for chemical applications. Statisticians, on the other hand, tend to be more focused on general methods that satisfy various mathematical optimality properties. The research of these communities has, however, overlapped from time to time.[16, 17] We have attempted to bridge the divide between the two communities by comparing the utility of five different chemometric and statistical techniques to the analysis and classification of ToF-SIMS spectral data. After analysis by PCA, commonly used by both chemometricians and statisticians, we have applied the common chemometric techniques of partial least squares discriminant analysis and soft independent modeling of class analogy and the statistical techniques of linear discriminant analysis and decision tree analysis for sample classification.

Principal Component Analysis (PCA) is an unsupervised multivariate technique that reduces a large data matrix to a few composite variables that can be visualized and interpreted using a series of simple plots. PCA reduces the data complexity by calculating new variables, called principal components, which represent linear combinations of the original variables and capture the greatest variation in the data set. PCA is a descriptive rather than classification technique; PCA does not classify samples into groups but rather helps a researcher to understand the relationships among sample groups and to identify variables important for explaining those differences between samples. While there have been several reports in the literature of PCA being adapted for classification ([18-20] among others), it was neither designed nor optimized for

this purpose.[21] We have therefore chosen to classify samples using a few of the statistical and chemometeric techniques specifically optimized for classification.[16, 22]

Linear discriminant analysis (LDA) is a supervised, multivariate statistical technique which specifically attempts to model the differences between classes in a data set.[22] LDA uses the known classes defined in the data set to calculate linear combinations of the original variables, called canonical variates, that maximize the ratio of the between-group variance to the within-group variance and are uncorrelated with each other. The LDA model thus created can then be used to predict the class membership of additional samples. The nature of the LDA calculations necessitates that the number of input variables be less than the number of samples in a given data set. In order to meet this requirement for mass spectral data, where the number of peaks typically far exceeds the number of samples, the data set must first be reduced by some method such as PCA prior to analysis by LDA.

Partial least squares discriminant analysis (PLSDA) is a supervised, chemometric classification technique which also utilizes known class information when creating new composite variables. In fact, Barker and Rayens[16] have shown that PLSDA is essentially an inverse-least squares approach to LDA and produces basically the same result, but with the advantage of performing the dimensional reduction along with classification in a single calculation. As with LDA, the PLSDA model can be used to predict the class membership of additional samples.

Soft independent modeling of class analogy (SIMCA) is a supervised, chemometric technique which models a data set with a collection of PCA models, one for each class in the data set. SIMCA is distinct from LDA and PLSDA in that each data class is modeled separately.

Additional samples can be assigned a class by calculating the nearest class to a sample, defined as the class model that results in a minimum distance of the sample to the mode.[23]

Use of a decision tree is another common statistical method to predict class information from a large, complex data set.[22] Decision trees, and the related technique of cluster analysis, are commonly used for analysis of large biological data sets. The construction of a decision tree is a fundamentally different type of analysis than those described above; rather than using linear combinations of variables to describe and classify the data set, a series of individual variables are chosen to create the tree and affect the classification.

In this work, we have analyzed ToF-SIMS spectral data from increasingly complex biological samples using a variety of both chemometric and statistical multivariate analyses. In order of increasing biological, and thus mass spectral, complexity, we have investigated monosaccharide isomers, pure proteins, protein mixtures, and mouse embryo tissues. The varying complexity of the sample sets has shown that with increasing sample complexity, one needs increasingly sophisticated multivariate analyses to classify samples effectively. The comparison of different chemometric and statistical techniques has shown that no one of these analyses is universally better at classifying these disparate types of samples. However, some types of multivariate analysis, namely PCA, PLSDA, and LDA, consistently perform better than others, including SIMCA and decision trees.

Experimental Section

Samples:

Monosaccharides: Experimental details of the monosaccharide analysis have been described elsewhere.[7] Briefly, galactose, glucose, fructose, mannose, psicose, sorbose, and tagatose were obtained from Sigma (St.Louis, MO) and used without further purification. Each

sugar was diluted in Milli-Q purified water (18.2 MΩ; Millipore, Billerica, MA) to a concentration of ~1 mg/mL, after which 1 μL was spotted on a silicon wafer and allowed to evaporate at room temperature. All seven sugars were spotted on each of seven 1.25 × 1.25 cm substrates. Ten individual spectra were acquired per spot.

Proteins: Myoglobin and cytochrome c from horse heart, lysozyme from chicken egg white, bovine insulin, and bovine albumin were obtained from Sigma and used without further purification. Each protein was diluted in Milli-Q purified water to a concentration of ~1 mg/mL. Individual proteins solutions (1 μL) were spotted randomly on five 1.25 × 1.25 cm substrates with sixteen protein spots per substrate, using sequences drawn from a uniform distribution on the unit interval. MatLAB v. R2006b (MathWorks Inc., Natick, MA) was used to generate the random sequences. Five individual spectra were recorded per spot.

Protein Mixtures: Glyceraldehyde-3-phosphate dehydrogenase from rabbit muscle, α-chymotrypsinogen A from bovine pancreas, and carbonic anhydrase were obtained from Sigma and thyroglobulin, aldolase, and ferritin were obtained from Pharmacia Biotech (Piscataway, NJ). All proteins were used without further purification. Five protein mixtures were created, each with a common, complex protein background and one distinct protein component per mixture. The common complex protein background consisted of equal concentrations of myoglobin, albumin, lysozyme, α-chymotrypsinogen A, and glyceraldehyde-3-phosphate dehydrogenase. One of cytochrome c, carbonic anhydrase, thyroglobulin, aldolase, or ferritin was added to this background for a total concentration of every protein in each mixture of approximately 0.16mg/mL and a total protein concentration of ~1 mg/mL. Randomized spotting was performed as described above. Five individual spectra were recorded per spot.

Mouse Embryos: Mouse embryo tissues were prepared as described elsewhere.[12] Briefly, three 16-day-old mouse embryos from three different dams were fixed in 4% paraformaldehye for 36 hours and embedded in paraffin blocks using standard techniques. Four-micron thick saggital sections were cut from each embryo using a Leica RM2165 microtome. The sections were placed on 1.25 × 1.25 cm silicon substrates and incubated at 40ºC overnight to soften the paraffin. The samples were then deparaffinized using xylene and 100% ethanol after which they were allowed to air dry. The samples were stored in vacuum at 1×10^{-4} Torr for 24 hours before ToF-SIMS analysis. Ten spectra were recorded for each tissue type for each embryo section.

ToF-SIMS Analysis: ToF-SIMS measurements were conducted on a PHI TRIFT III mass spectrometer (Physical Electronics USA, Chanhassen, MN). Sugar, protein and mouse embryo data were acquired with a gold ($^{197}Au^{+}$) liquid metal ion gun operated at 22 kV while protein mixture data were acquired with a gallium ($^{69}Ga^{+}$) liquid metal ion gun operated at 15 kV. Positive ion ToF-SIMS spectra were acquired over an area of 100 × 100 μm for sugar, protein and protein mixture samples. Positive ion ToF-SIMS images were acquired over an area of 300 × 300 μm for mouse embryo tissues, after which one region of interest spectrum was extracted for the specific tissue type of interest. ToF-SIMS spectra were calibrated to the CH_3^+, $C_2H_3^+$, and $C_4H_7^+$ peaks before further analysis.

Statistical and Chemometric Analyses:

Preprocessing: Each spectral data set was minimally preprocessed prior to multivariate analysis by unit-mass binning of peaks, choosing an appropriate peak set for the data, normalizing to the total ion count of the chosen peaks, and mean centering. For the sugar data, peaks between *m/z* = 12 and 500 were utilized with the exception of peaks due to sodium (*m/z* =

23), potassium (*m/z* =39), and the silicon substrate (*m/z* = 28). For the proteins and protein mixtures, peaks between *m/z* = 60 and 500 were utilized in order to exclude non-specific hydrocarbon peaks with up to four carbons. In addition, peaks due to PDMS contamination (*m/z* = 73, 147, 207, 221, and 281) were removed. For the mouse embryo samples, the calcium peak (*m/z* = 40) and peaks between *m/z* = 60 and 500 were utilized with the exception of those contamination peaks identified from the background control spectra as described elsewhere.[12] Although there have been several recent publications on more sophisticated methods for preprocessing ToF-SIMS data,[24-28] we chose simple and widely-used preprocessing steps for this study. A thorough a comparison of preprocessing methods is beyond the scope of this current work.

Multivariate Analyses: The five different chemometric and statistical multivariate analyses were run using two standard software packages. MatLAB software v. R2006b (MathWorks Inc., Natick, MA) along with PLS Toolbox v. 4.1 (Eigenvector Research, Manson, WA) was used for PCA, PLSDA and SIMCA. The R Environment for Statistical Computing[29] was utilized for PCA, LDA, and a decision tree analysis (using the rpart package). PCA was performed using both software packages and in all cases identical results were obtained.

Cross-validation: For any classification method one desires to know its generalizability, that is, not just how well it classifies the data with which it was trained, but how well it would do on a new data set. We used cross-validation to estimate this generalization error for each of the classification methods studied. Specifically, for each method, we repeatedly extracted a random subset of the data to use as a test set and used the remaining data as a training set for reducing the data and building the classifier. After creating each classifier using the associated random training set, we then used the test set to estimate the misclassification rate. The average

misclassification rate for 100 random test sets was reported as the estimated misclassification rate, and the standard deviation over the 100 test sets was reported as the standard error in estimating the misclassification rate. We chose a test set size consisting of the nearest integer to the square root of the total data set size. This test set size represents a compromise between choosing a test set size large enough to reduce the standard error in estimating the misclassification rate, while keeping the training set size large enough to build an accurate classifier.

Linear Mixed Effects Modeling: To better understand the impact of variation arising from different spots of the same sample and different substrates, we have estimated the between-substrate standard deviation and the within-substrate standard deviation by means of a linear mixed-effect model.[30] The model used was the standard blocked one-way layout, containing a fixed effect for the samples and two nested random effects: one for sample substrates and one for the residual deviation within any given substrate. This analysis was performed for both the sugar and protein data sets. Linear mixed effects modeling showed a non-negligible contribution to the residual deviation for both random effects. To alleviate the impact of these effects, the placement of the sample spots on the silicon substrates was randomized as described for the protein and protein mixture experiments. Due to the nature of the mouse embryo tissue samples, it was not possible to randomize the placement on the substrates, but careful analysis of the PCA results showed no obvious grouping due to substrate or sample effects.

Results and Discussion

The nature of the spectral data produced by ToF-SIMS analysis is illustrated in Figure 1, which shows representative spectra from each of the four data sets. These four spectra

demonstrate the high degree of fragmentation produced by the SIMS ionization, as is evident from the large number of low molecular mass peaks. In addition, the spectra from cytochrome c, a complex protein mixture, and a mouse rib (Figure 1 b,c,d) show the inherent complexity and similarity of the mass spectra derived from such complex biological samples.

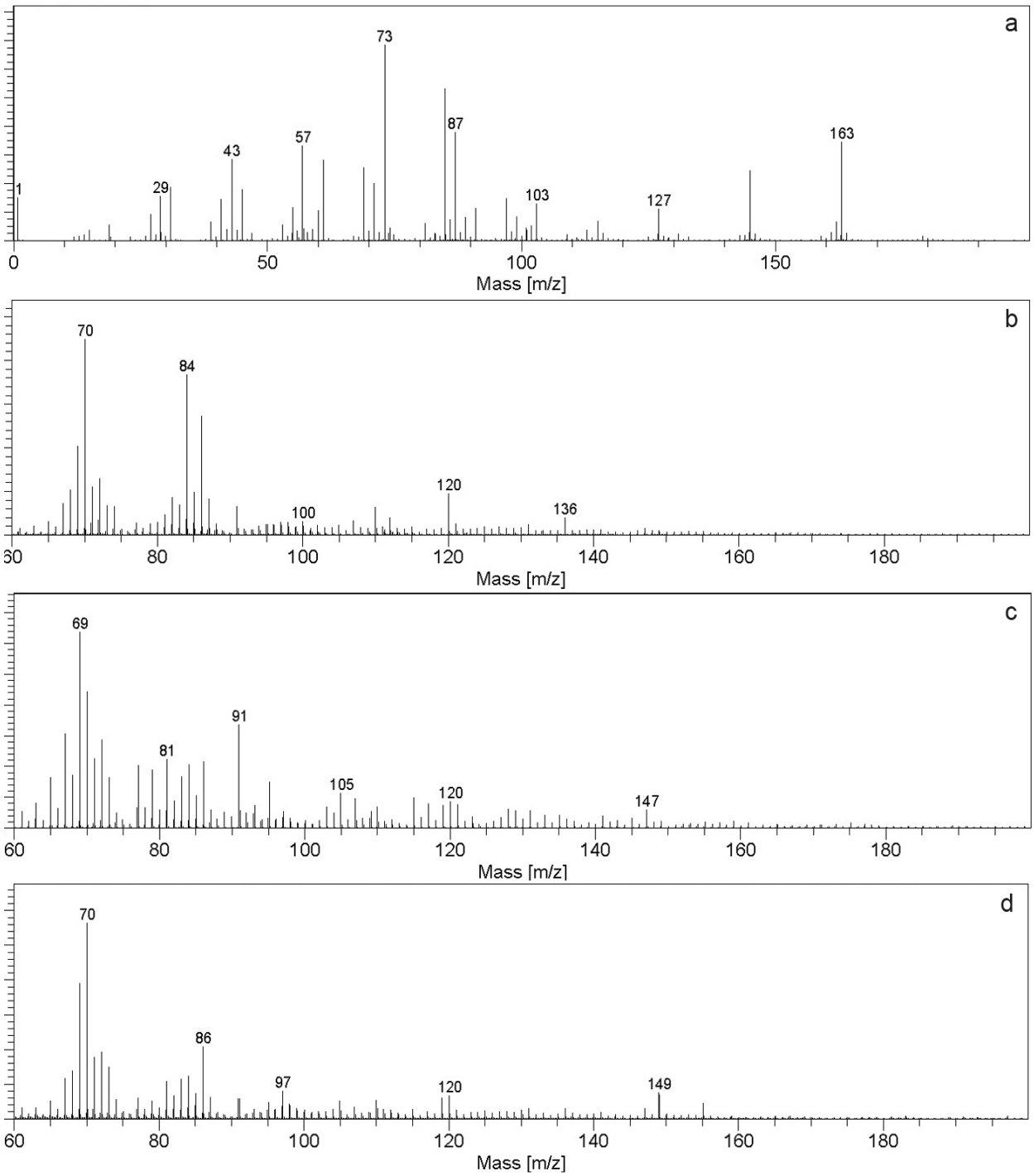

Figure 1. Representative ToF-SIMS spectra from the four data sets used for statistical analysis: a. glucose; b. pure cytochrome c; c. a mixture of albumin, myoglobin, lysosyme, α-chymotrypsinogen A, glyceraldehyde -3-phosphate, and cytochrome c; and d. a rib from a mouse embryo.

As is apparent from Figure 1, it is generally not possible by visual inspection of ToF-SIMS spectra of complex samples to classify a spectrum as arising from a particular sample of interest

or even to identify completely the molecular fragments which vary most significantly between samples. It is therefore necessary to utilize a suitable statistical analysis technique to gain these types of insights from the ToF-SIMS spectra.

As an example of a chemically challenging but biologically simple sample set, we recorded spectra from seven monosaccharide sugar isomers, galactose, glucose, fructose, mannose, psicose, sorbose, and tagatose. The resulting well-controlled data set was analyzed using five different chemometric and multivariate statistical analyses. As has been shown in previous work,[7] PCA of these seven sugar isomers shows excellent grouping of spectra by isomer. Based on the results of this unsupervised multivariate analysis, it is not surprising that the three supervised classification techniques, LDA, PLSDA, and SIMCA, are easily able to classify sugars by isomer with zero misclassification rate given a sufficient number of principal components (Figure 2). Note that while the compound variables created by LDA and PLSDA are often referred to as latent variables, we have chosen to use the term "principal component" for all the analysis techniques for the sake of simplicity.

Cursory inspection of Figure 2 would seem to indicate that SIMCA is the best method for classifying the samples at low numbers of principal components. However, because a SIMCA analysis calculates a separate PCA model for each sample class, one principal component in a SIMCA model is actually one for each class, or in this case 7 compound variables. At seven principal components, both LDA and PLSDA are able to classify the sugar samples with no misclassification. Based on our analysis, LDA following data reduction by PCA is able to perfectly classify the sugar samples using only four compound variables, making it the best of these methods for classifying the spectra of these sugar isomers. It is interesting to note, given the similarity between the calculations performed for LDA and PLSDA,[16] that LDA performs

appreciably better in classifying this data set with an intermediate number of principal components than does PLSDA. This is probably due to the fact that, as Barker and Rayens note, with PLSDA "there is no claim to minimizing misclassification probabilities" as there is for LDA.[16]

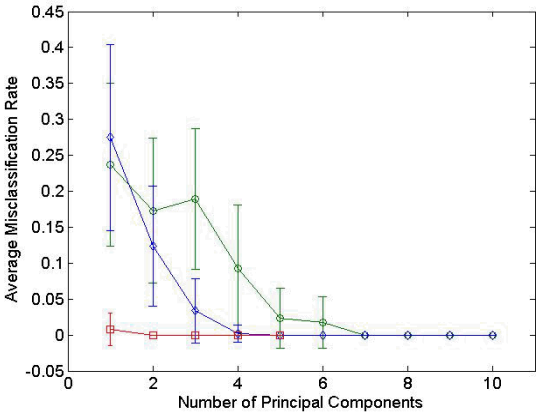

Figure 2. Comparison of average rates of misclassification for ToF-SIMS spectra of seven monosaccharides using three different classification methods; LDA following PCA (blue diamonds), PLSDA (green circles), and SIMCA (red squares). Error bars represent one standard deviation.

While LDA, PLSDA, and SIMCA all give perfect classification results for the sugar isomer data set, the decision tree algorithm is less successful. The decision tree produces an average misclassification rate of 0.0207 with a standard deviation of 0.0383 for 100-fold cross-validation. Perhaps more importantly, inspection of the tree produced by the recursive partitioning algorithm shows that this analysis provides less chemical insight than we had hoped (Figure 3). First, we would expect that the sugars would be grouped according to their known chemical structures, with primary separation into pyranose and furanose groups. Instead, we see that sorbose is the first sugar to be grouped, while the remaining furanoses, fructose, psicose, and tagatose, are grouped together at the bottom of the tree. Second, the variables selected as the decision points are not, in many cases,

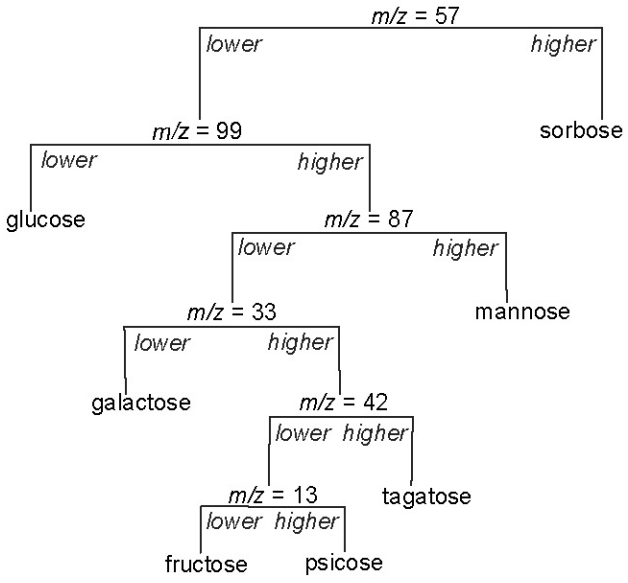

Figure 3. Decision tree generated by recursive partitioning to classify sugar isomers based on their ToF-SIMS spectra.

either the most abundant fragments or even fragments specific to these monosaccharides, as demonstrated by the CH^+ fragment (*m/z*=13) which defines the classification of fructose and psicose. Interestingly, analysis by PCA has neither of these drawbacks: there is obvious separation into pyranose and furanose groups and the molecular fragments with the highest loadings on the first few principal components are identified sugar fragments.[7] It is clear that in the case of the ToF-SIMS analysis of sugar isomers, the decision tree analysis is the least successful of the attempted methods, both in terms of classification and data interpretation.

We have applied these same data analysis techniques to spectra from five different pure proteins, samples which are more chemically different and of increasing biological complexity. Figure 4a shows the scores plot producing the best grouping of proteins from a PC analysis of spectra from myoglobin, cytochrome c, lysozyme, insulin, and albumin. PCA is clearly able to distinguish among the ToF-SIMS spectra of these protein standards. Not surprisingly, classifications by LDA, PLSDA, and SIMCA again produce excellent results with essentially zero misclassifications with 8 or more principal components. It is notable that while more principal components are necessary for perfect classification with LDA and PLSDA than were needed for the sugar classification, the misclassification rate is lower for both of these methods at low numbers of principal components. It is also the case that the SIMCA analysis performs significantly worse on the proteins, with larger misclassification rates and much larger standard deviations. Once again, LDA following data reduction by PCA appears to be the best of these methods for classifying the spectra. The decision tree analysis of the protein data set produces an average misclassification rate of 0.023 with a standard deviation of 0.0351 for 100-fold cross-validation. As with the analysis of sugars, the decision tree analysis was less accurate for

classification and provided no additional chemical insight. (The decision tree is available in supporting information.)

While sugars and proteins are interesting biological molecules, we are primarily interested in understanding how statistical and chemometric analyses will be best applied to ToF-SIMS analysis of much more complex, native biological samples. In order to create a well-controlled sample set which closely mimics the complexities of biological fluids, cells, and tissues, we created a set of protein mixtures such that each mixture contains a common, complex protein background and one distinct protein component. These mixtures are described in detail in the experimental section. Figure 5 shows the results of the multivariate analyses of ToF-SIMS spectra of these complex protein mixtures.

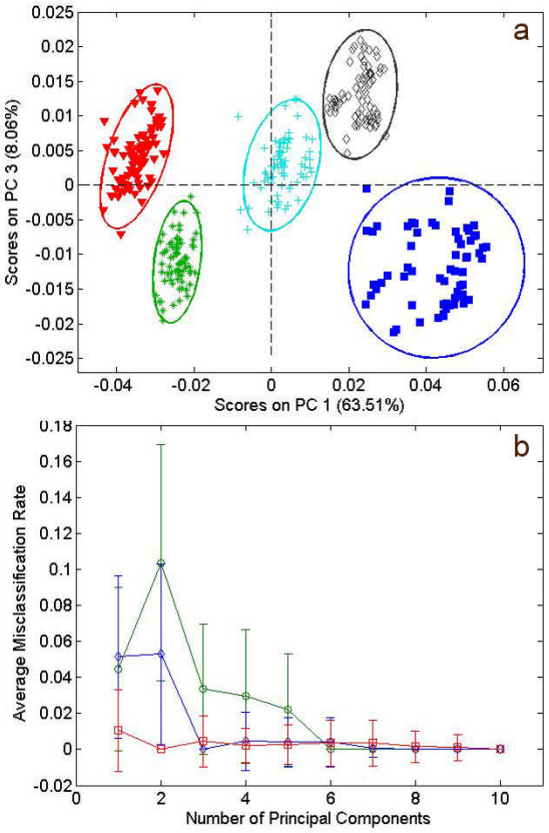

Figure 4. Multivariate analyses of ToF-SIMS spectra of five pure proteins spotted on silicon substrates: a. Scores plot from PCA data reduction; myoglobin (red triangles), cytochrome c (green stars), lysosyme (blue squares), albumin (cyan plusses) and catalase (black diamonds). Each point is a single spectrum. Ellipses are 95% confidence ellipses. b. Comparison of average rates of misclassification using three different classification methods; LDA following PCA (blue diamonds), PLSDA (green circles), and SIMCA (red squares). Error bars represent one standard deviation.

It is obvious from the best grouping obtained by principal component analysis (Figure 5a) that this set of complex protein mixtures is much more difficult to differentiate than are the pure sugars or proteins. Nevertheless, PCA does provide some insight into the data, dividing the samples into distinct groupings, with mixtures containing aldolase and ferritin obviously

different from the other three. The mixture containing thyroglobulin is also reasonably well separated from those containing cytochrome c and carbonic anhydrase. The PC analysis alone, however, is clearly insufficient for a complete analysis of this protein mixture data.

Figure 5b details the analysis of the protein mixture samples using LDA, PLSDA, and SIMCA. It is immediately obvious from the very high misclassification rates at lower numbers of principal components, and from the large numbers of principal components needed for good classification, that this data set is inherently more complex and difficult to analyze. However, even with the high sample complexity, misclassification rates of less than 2% can be achieved with a sufficient number of principal components. In contrast to the pure sugars and proteins, PLSDA produces better classification results for the protein mixtures with any number of principal components. The reason for this shift may be attributable to an increase in the similarity of the spectra in this case.[16] The decision tree analysis of the protein mixture data produces an average misclassification rate of

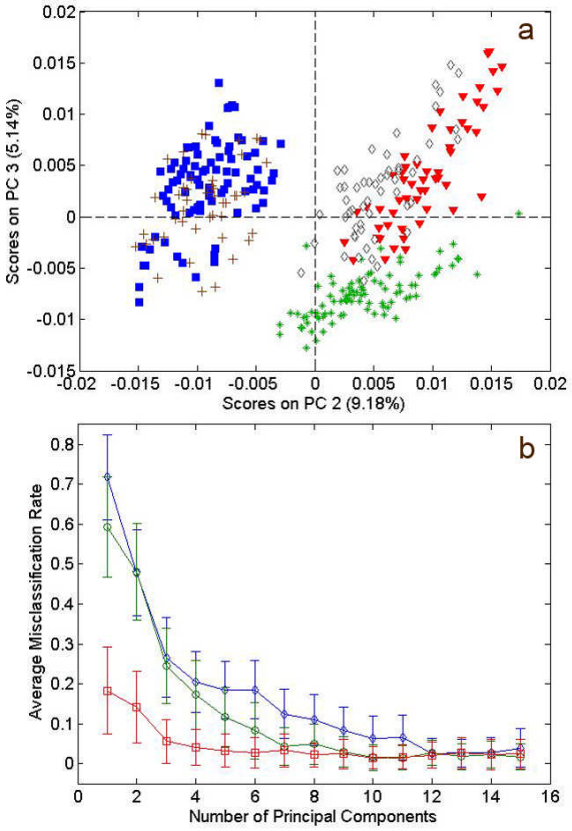

Figure 5. Multivariate analyses of ToF-SIMS spectra of five protein mixtures spotted on silicon substrates. a. Scores plot from PCA data reduction; Each mixture contains equal concentrations of albumin, myoglobin, lysosyme, α-chymotrypsinogen A and glyceraldehyde -3-phosphate in addition to one unique protein; cytochrome c (red triangles), thyroglobulin (green stars), aldolase (blue squares), ferritin (brown plusses) or carbonic anhydrase (black diamonds). Each point is a single spectrum. b. Comparison of average rates of misclassification using three different classification methods; LDA following PCA (blue diamonds), PLSDA (green circles), and SIMCA (red squares). Error bars represent one standard deviation.

0.0756 with a standard deviation of 0.0739 for 100-fold cross-validation. As for the other data sets, the decision tree analysis was less accurate for classification and provided no additional chemical insight. (The decision tree is available in supporting information.)

The final sample set we have analyzed consists of formalin-fixed, paraffin embedded mouse embryo tissues. These samples represent the full complexity of the types of biological samples one would like to be able to study utilizing ToF-SIMS and multivariate analysis. The results from PC analysis were reported previously,[12] in which PCA of four tissue types, rib, brain, heart, and liver, shows good grouping of spectra by tissue. Notably, PCA demonstrated better differentiation of the four different tissue types than of the protein mixture standards. This result is not surprising given the dissimilar types of tissues chosen for this analysis and the purposeful extreme similarity of the protein mixtures.

The three supervised multivariate techniques also show that the mouse embryo tissues are less challenging to classify than the protein mixtures but more challenging than the pure proteins or sugars (Figure 6). PLSDA and LDA perform equally well at classifying these tissue samples, with both methods producing less than 2% misclassification with a sufficient number of principal components. SIMCA, however, performs much more poorly on the mouse data set, with misclassification rates around 5% for low numbers of principal components and increasing misclassification with increasing numbers of principal components. It is not

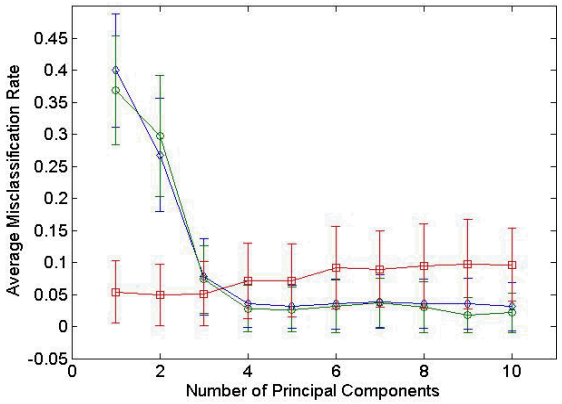

Figure 6. Comparison of average rates of misclassification for ToF-SIMS spectra of four tissues from paraffin-embedded mouse embryos using three different classification methods; LDA following PCA (blue diamonds), PLSDA (green circles), and SIMCA (red squares). Error bars represent one standard deviation.

immediately obvious why the SIMCA results are so different for the mouse embryo tissues. The decision tree analysis of the mouse embryo tissues produces an average misclassification rate of 0.0776 with a standard deviation of 0.0634 for 100-fold cross-validation. As for the other data sets, the decision tree analysis was less accurate for classification and provided no additional chemical insight. (The decision tree is available in supporting information.)

Conclusions

We conclude that principal component analysis is an excellent first step in examining ToF-SIMS spectral data of biological samples. PCA is straightforward and quick to implement, easily understood, and produces a series of simple plots for examining the data. Furthermore, PCA provides insight both into the similarities and differences among sample groups and into the mass spectral peaks which are most important for determining group differences. These insights can ultimately be utilized to gain a molecular understanding of how samples differ. PCA is, however, insufficient for a thorough analysis of all but the simplest sample sets. As we have shown, the ability of PCA to distinguish among sample groups diminishes as the similarity of the groups and the complexity of the samples increases. Furthermore, PCA does not provide for a statistically rigorous classification of future unknown samples into sample classes.

Among the four classification methods utilized in this study, the decision tree analysis is clearly the least successful for analysis of ToF-SIMS spectra. The rate of misclassification was relatively high when compared to the other methods, groupings produced by the recursive partitioning did not organize samples according to known sample similarities, and the variables selected as the decision points were not, in most cases, either the most abundant fragments or fragments specific to the samples of interest. The poor results obtained by the decision tree

analysis most likely arise from the fact that this analysis uses a series of univariate decision points, rather than a decision based on a multivariate set of compound variables, thus utilizing only a small subset of the information contained in the mass spectral data.

Classification analysis by SIMCA was also less successful than some of the other methods. As SIMCA is essentially multiple applications of PCA, it should not be surprising that it shares some of the disadvantages of PCA for analysis of more complex samples. Specifically, the ability of SIMCA to accurately classify samples decreased as the complexity of the samples increased. Therefore it is concluded that SIMCA is not well suited for analyzing spectral data sets from native biological samples,

In contrast, both PLSDA and LDA following data reduction by PCA produced excellent classification results for all four data sets examined. While the misclassification rates did increase for the most complex samples, rates of less than 2% misclassification were achieved. Given the underlying similarities of these two methods, it is somewhat surprising that the results were not always equivalent, especially with an intermediate number of compound variables calculated. However, the best classification rate achieved with the two methods was in every case very similar. From this study, we conclude that, if computation resources allow, attempting both PLSDA and LDA and comparing the results would be ideal. However, the choice of only one of these two would reduce analysis time while not unduly reducing the likelihood of excellent classification results. We further note that using these multivariate analysis techniques to analyze ToF-SIMS spectra is not in any way specific to the analysis of biological samples and should also be well suited to the analysis of other complex sample systems such as inorganic and synthetic polymer samples.

As the complexity of the samples increases, so must the sophistication of the chemometric or statistical analysis methods employed to understand and classify the ToF-SIMS spectra. It is not possible, therefore, to designate one multivariate analysis technique as the "best" for analysis of ToF-SIMS spectra of all biological samples; rather it will be necessary to evaluate a few appropriate techniques for each application. This study demonstrates that ToF-SIMS spectral analysis in conjunction with common statistical and chemometric techniques can be effectively utilized to classify complex biological samples and opens the possibilities for new applications including classification of subtly different biological samples that may provide insights into cellular processes, disease progress, and disease diagnosis.

Acknowledgement

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and supported by CA BCRP 11NB-0178 and LDRD 04-ERD-104 (LLNL internal funding).

Supporting Information Available: Additional information as noted in the text.

References

(1) Brentani, R. R.; Carraro, D. M.; Verjovski-Almeida, S.; Reis, E. M.; Neves, E. J.; de Souza, S. J.; Carvalho, A. F.; Brentani, H.; Reis, L. F. L. *Critical Reviews in Oncology/Hematology* **2005**, *54*, 95-105.
(2) Macgregor, P. F. *Expert Review of Molecular Diagnostics* **2003**, *3*, 185-200.
(3) Belu, A. M.; Graham, D. J.; Castner, D. G. *Biomaterials* **2003**, *24*, 3635-3653.
(4) Lockyer, N. P.; Vickerman, J. C. *Applied Surface Science* **2004**, *231-232*, 377-384.
(5) Michel, R.; Castner, D. G. *Surface and Interface Analysis* **2006**, *38*, 1386-1392.
(6) Johansson, B. *Surface and Interface Analysis* **2006**, *38*, 1401-1412.
(7) Berman, E. S. F.; Kulp, K. S.; Knize, M. G.; Wu, L.; Nelson, E. J.; Nelson, D. O.; Wu, K. J. *Analytical Chemistry* **2006**, *78*, 6497-6503.

(8) Kulp, K. S.; Berman, E. S. F.; Knize, M. G.; Shattuck, D. L.; Nelson, E. J.; Wu, L.; Montgomery, J. L.; Felton, J. S.; Wu, K. J. *Analytical Chemistry* **2006**, *78*, 3651-3658.
(9) Jungnickel, H.; Jones, E. A.; Lockyer, N. P.; Oliver, S. G.; Stephens, G. M.; Vickerman, J. C. *Analytical Chemistry* **2005**, *77*, 1740-1745.
(10) Fletcher, J. S.; Henderson, A.; Jarvis, R. M.; Lockyer, N. P.; Vickerman, J. C.; Goodacre, R. *Applied Surface Science* **2006**, *252*, 6869-6874.
(11) Thompson, C. E.; Ellis, J.; Fletcher, J. S.; Goodacre, R.; Henderson, A.; Lockyer, N. P.; Vickerman, J. C. *Applied Surface Science* **2006**, *252*, 6719-6722.
(12) Wu, L.; Lu, X.; Kulp, K. S.; Knize, M. G.; Berman, E. S. F.; Nelson, E. J.; Felton, J. S.; Wu, K. J. *International Journal of Mass Spectrometry* **2007**, *260*, 137-145.
(13) Gallagher, N. B.; Shaver, J. M.; Martin, E. B.; Morris, J.; Wise, B. M.; Windig, W. *Chemometrics and Intelligent Laboratory Systems* **2004**, *73*, 105-117.
(14) Graham, D. J.; Wagner, M. S.; Castner, D. G. *Applied Surface Science* **2006**, *252*, 6860-6868.
(15) Klerk, L. A.; Broersen, A.; Fletcher, I. W.; van Liere, R.; Heeren, R. M. A. *International Journal of Mass Spectrometry* **2007**, *260*, 222-236.
(16) Barker, M.; Rayens, W. *Journal of Chemometrics* **2003**, *17*, 166-173.
(17) Ding, B.; Gentleman, R. *Journal of Computational and Graphical Statistics* **2005**, *14*, 280-298.
(18) Ingram, J. C.; Bauer, W. F.; Lehman, R. M.; O'Connell, S. P.; Shaw, A. D. *Journal of Microbiological Methods* **2003**, *53*, 295-307.
(19) Sanni, O. D.; Wagner, M. S.; Briggs, D.; Castner, D. G.; Vickerman, J. C. *Surface and Interface Analysis* **2002**, *33*, 715-728.
(20) Wagner, M. S.; Castner, D. G. *Langmuir* **2001**, *17*, 4649-4660.
(21) Jackson, J. E. *A User's Guide to Principal Components*; Wiley-Interscience: Hoboken, NJ, 2003.
(22) Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S*, 4 ed.; Springer: New York, 2002.
(23) Wold, S. *Pattern Recognition* **1976**, *8*, 127-139.
(24) Keenan, M. R.; Kotula, P. G. *Surface and Interface Analysis* **2004**, *36*, 203-212.
(25) Keenan, M. R.; Kotula, P. G. *Applied Surface Science* **2004**, *231-232*, 240-244.
(26) Wagner, M. S.; Graham, D. J.; Ratner, B. D.; Castner, D. G. *Surface Science* **2004**, *570*, 78-97.
(27) Tyler, B. J. *Applied Surface Science* **2003**, *203-204*, 825-831.
(28) Tyler, B. J.; Rayal, G.; Castner, D. G. *Biomaterials* **2007**, *28*, 2412-2423.
(29) Team, R. D. C., 2.4.1 ed.; R Foundation for Statistical Computing: Vienna, Austria, 2006.
(30) Pinheiro, J. C.; Bates, D. M. *Mixed Effects Models in S and S-PLUS*; Springer-Verlag: New York, 2000.